# Method and System for Detecting and Temporally Relating Components in Non-Stationary Signals

## Field of the Invention

[01] The invention relates generally to the field of signal processing and in particular to detecting and relating components of signals.

## Background of the Invention

[02] Detecting components of signals is a fundamental objective of signal processing. Detected components of acoustic signals can be used for myriad purposes, including speech detection and recognition, background noise subtraction, and music transcription, to name a few. Most prior art acoustic signal representation methods have focused on human speech and music where detected component is usually a phoneme or a musical note. Many computer vision applications detect components of videos. Detected components can be used for object detection, recognition and tracking.

[03] There are two major types of approaches to detecting components in signals, namely knowledge based, and unsupervised or data driven. Knowledge-based approaches can be rule-based. Rule-based approaches require a set of human-determined rules by which decisions are made. Rule-based component detection is therefore subjective, and decisions on occurrences of components are not based on actual data to be analyzed. Knowledge based system have serious disadvantages. First, the rules need to

be coded manually. Therefore, the system is only as good as the 'expert'. Second, the interpretation of inferences between the rules often behaves erratically, particularly when there is no applicable rule for some specific situation, or when the rules are 'fuzzy'. This can cause the system to operate in an unintended and erratic manner.

[04] The other major types of approach to detecting components in signals are data driven. In data driven approaches, the components are detected directly from the signal itself, without any a priori understanding of what the signal is, or could be in the future. Since input data is often very complex, various types of transformations and decompositions are known to simplify the data for the purpose of analysis.

[05] U.S. Patent 6,321,200, "Method for extracting features from a mixture of signals," issued to Casey on November 20, 2001 describes a system that extracts low level features from an acoustic signal that has been band-pass filtered and simplified by a singular value decomposition. However, some features cannot be detected after dimensionality reduction because the matrix elements lead to cancellations, and obfuscate the results.

[06] Non-negative matrix factorization (NMF) is an alternative technique for dimensionality reduction, see, Lee, et al, "Learning the parts of objects by non-negative matrix factorization," Nature, Volume 401, pp.788-791, 1999.

[07] There, non-negativity constraints are enforced during matrix construction in order to determine parts of faces from a single image. Furthermore, that system is restricted within the spatial confines of a single image, that is, the signal is stationary.

## Summary of the Invention

[08] The invention provides a method for detecting components of a non-stationary signal. The non-stationary signal is acquired and a non-negative matrix of the non-stationary signal is constructed. The matrix includes columns representing features of the non-stationary signal at different instances in time. The non-negative matrix is factored into characteristic profiles and temporal profiles.

## Brief Description of the Drawings

[09] Figure 1 is a block diagram of a system for detecting non-stationary signal components according to the invention;

[010] Figure 2 is a flow diagram of a method for detecting non-stationary signal components according to the invention;

[011] Figure 3 is a spectrogram to be represented as a non-negative matrix;

[012] Figure 4A is a diagram of temporal profiles of the spectrogram of Figure 3;

[013] Figure 4B is a diagram of characteristic profiles of the spectrogram of Figure 3;

[014] Figure 5 is a bar of music with a temporal sequence of notes;

[015] Figure 6 is a block diagram correlating the profiles of Figures 4A-4B with the bar of music of Figure 5;

[016] Figure 7A is a temporal profile;

[017] Figure 7B is a characteristic profile;

[018] Figure 8 is a block diagram of a video with a temporal sequence of frames;

[019] Figure 9A is a temporal profile of the video of Figure 8;

[020] Figure 9B is a characteristic profile of the video of Figure 8; and

[021] Figure 10 is a schematic of a piano action.

## Detailed Description of the Preferred Embodiment

### Introduction

[022] As shown in Figures 1 and 2, the invention provides a system 100 and method 200 for detecting components of non-stationary signals, and determining a temporal relationship among the components.

### System Structure

[023] The system 100 includes a sensor 110, e.g., microphone, an analog-to-digital (A/D) converter 120, a sample buffer 130, a transform 140, a matrix buffer 150, and a factorer 160, serially connected to each other. An acquired non-stationary signal 111 is input to the A/D converter 120, which outputs samples 121 to the sample buffer 130. The samples are windowed to produce frames 131 for the transform 140, which outputs features 141, e.g., magnitude spectra, to the matrix buffer 150. A non-negative matrix 151 is factored 160 to produce characteristic profiles 161 and temporal profiles 162, which are also non-negative matrices.

### Method Operation

[024] An acoustic signal 102 is generated by a piano 101. The acoustic signal is acquired 210, e.g., by the microphone 110. The acquired signal 111 is sampled and converted 220 and digitized samples 121 are windowed 230. A transform 140 is applied 240 to each frame 131 to produce the features 141. The features 141 are used to construct 250 a non-negative matrix 151.

The matrix 151 is factored 260 into the characteristic profiles 161 and the temporal profiles 162 of the signal 102.

**[025] Constructing the Non-Negative Matrix**

[026] An example of the time-varying signal 102 can be expressed by $s(t) = g(\alpha t)\sin(\gamma t) + g(\beta t)\sin(\delta t)$, where $g(\cdot)$ is a gate function with a period of $2\pi$ and $\alpha, \beta, \gamma, \delta$ are arbitrary scalars with $\alpha$ and $\beta$ at least an order of magnitude smaller than $\gamma$ and $\delta$. The features 141 of the frames $\mathbf{x}(t)$ 131, having a length size $L$, are determined by a transform $\mathbf{x}(t) = |\text{DFT}([s(t) \ldots s(t+L)])|$ 140.

[027] The non-negative matrix $\mathbf{F} \in \mathrm{R}^{M \times N}$ 151 is constructed 250 by arranging all the features 141 as $N$ columns of the matrix 151 ordered temporally with $M$ rows, where $M$ is the total number of histogram bins into which the magnitude spectra features are accumulated, such that $M = (L/2+1)$.

[028] Figure 3 shows a binned spectrogram to be represented as the non-negative matrix 151 $\mathbf{F}$ of the signal $s(t)$. This example has little energy except for a few frequency bins 310. The bins display a regular pattern.

**[029] Non-Negative Matrix Factorization**

[030] As shown in Figures 4A-4B, the non-negative matrix $\mathbf{F} \in \mathrm{R}^{M \times N}$ is factored into two non-negative matrices $\mathbf{W} \in \mathrm{R}^{M \times R}$ (161) and $\mathbf{H} \in \mathrm{R}^{R \times N}$

(162), where $R \leq M$, such that an error in a non-negative matrix reconstructed from the factors is minimized.

[031] The parameter $R$ is the desired number of components to be detected. If the actual number of components in the signal is known, parameter $R$ is set to that known number and the error of reconstruction is minimized by minimizing a cost function $C = \| \mathbf{F} - \mathbf{W} \cdot \mathbf{H} \|_F$ where $\| \cdot \|_F$ is the Frobenius norm. Alternatively, if $R$ is set to an estimate of the number of components, then the cost function can be minimized by

$$D = \left\| \mathbf{F} \otimes \ln\left( \frac{\mathbf{F}}{\mathbf{W} \cdot \mathbf{H}} \right) - \mathbf{F} + \mathbf{W} \cdot \mathbf{H} \right\|_F ,$$

[032] where $\otimes$ is a Hadamard product. Both $C$ and $D$ equal zero if $\mathbf{F} = \mathbf{W} \cdot \mathbf{H}$.

[033] Figures 4B and 4A show respectively the spectral profiles 161 and the characteristic profiles 162 produced by the NMF on the matrix 151. In this case, the characteristic profiles of the components relate to frequency features. It is clear that component 1 occurs twice, and component 2 occurs thrice, compare with Figure 3.

**[034] Results**

[035] The system and method according to the invention was applied to a piano recording of Bach's fugue XVI in G minor, see Jarrett, "J.S. Bach, Das Wohltemperierte Klavier, Buch I", *ECM Records*, CD 2, Track 8, 1988.

Figures 5 shows one bar 501 of four distinct notes, with one note repeated twice. The recording was sampled at a rate of 44,100kHz and converted to a monophonic signal by averaging the left and right channels of the stereophonic signal. The samples were windowed using a Hanning window. A 4096-point discrete Fourier transform was applied to each frame to generate the columns of the non-negative matrix. The first matrix was factored using the first cost function for $R = 4$.

[036] Figure 6 shows a correlation between the profiles and the bar of notes.

[037] Figure 7 show profiles produced by the factorization when the parameter R is 5, and the second cost function is used. The extra temporal profiles 701 can be identified by their low energy wideband spectrum. These profiles do not correspond to any components, and can be ignored.

**[038] Constructing a Non-Negative Matrix for Analysis of Video**

[039] The invention is not limited to 1D linear acoustic signal. Components can also be detected in non-stationary signals with higher dimensions, for example 2D. In this case, the piano 101 remains the same. The signal 102 is now visual, and the sensor 110 is a camera that converts the visual signal to pixels, which are sampled, over time, into frames 131, having an area size ($X$, $Y$). The frames can be transformed 140 in a number of ways, for example by rasterization, FFT, DCT, DFT, filtering, and so

forth depending on the desired features to characterize for detection and correlation, e.g., intensity, color, texture, and motion.

[040] Figure 8 shows 2D frames 800 of a video. This action video has two simple components (rectangle and oval), each blinking on and off. In this example, the *M* pixels in each of the *N* frame are rasterized to construct the columns of the non-negative matrix 151.

[041] Figures 9A-9B show the characteristic profiles 161 and the temporal profiles 162 of the components of the video, respectively. In this case, the characteristic profiles of the components relate to spatial features of the frames.

[042] As a further example, to illustrate the generality of the invention, the non-stationary signal can be in 3D. Again, the piano remains the same, but now one peers inside. The sensor is a scanner, and the frames become volumes. Transformations are applied, and profiles 161-162 can be correlated.

[043] It should be noted that the 1D acoustic signal, 2D visual signal, and 3D scanned profiles can also be correlated with each other when the acoustic, visual, and scanned signals are acquired simultaneously, since all of the signals are time aligned. Therefore, the motion of the piano player's fingers can, perhaps, be related to the keys as they are struck, rocking the rail, raising the sticker and whippen to push the jack heel and hammer, engaging the spoon and damper, until the action 1000 causes the strings to vibrate to produce the notes, see Figure 10.

[044] Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications may be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.